DOCUMENT RESUME

ED 282 416 FL 016 707

AUTHOR Theunissen, Phiel J. J. M.
TITLE Introduction to Rasch Measurement: Some Implications for Languages.
PUB DATE May 83
NOTE 5p.; In: van Weeren, J., Ed. Practice and Problems in Language Testing 5; see FL 016 706.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability Grouping; Difficulty Level; Evaluation Criteria; *Item Sampling; *Language Tests; *Measurement Techniques; Performance Factors; Second Languages; *Statistical Analysis; Testing Problems; *Test Items; Test Theory; *Test Validity
IDENTIFIERS *Rasch Model

ABSTRACT

Any systematic approach to the assessment of students' ability implies the use of a model. The more explicit the model is, the more its users know about what they are doing and what the consequences are. The Rasch model is a strong model where measurement is a bonus of the model itself. It is based on four ideas: (1) separation of observable behavior and an underlying latent trait responsible for that behavior; (2) the probabilistic character of the observed behavior; (3) local stochastic independence; and (4) repeated assessment of indices of the latent trait. The Rasch model specifies that the student's performance is completely determined by two variables, his/her ability and the item difficulty, to the exclusion of all other factors such as time, guessing, and the discriminative power of the items. The model has these characteristics: it has a theory of behavior before measurement, in contrast with classical test theory; it is geared to the item level rather than the test level; it does not try to explain why a certain answer is given, only the chance of a correct response; it is population-independent; and it permits clearer insight into the accuracy of the measurement. The model can be used for generating items from an item bank for different populations or different purposes. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED282416

In stark contrast with the well-known 'boy-meets-girl' situation, surprisingly little is known about the 'pupil-meets-test' situation. As test developers, we know at least one thing for sure, that is, we have to measure the pupils ability or performance, and one way to do this is by using the Rasch-model. In this one sentence I have used two concepts, the understanding of which is important to an understanding of test-theory. These are the concepts of 1. measurement and 2. model.

What is measurement? A very simple definition is: 'measurement is the assignment of numbers to object people according to certain rules.' Examples of rules are the ways in which we measure length (e.g. using an object calibrated with a standard in Paris) or temperature in centrigrades. We find it then completely 'normal' that we can say things like: "This stick is twice as long as that stick, or 10° C is equivalent to 50° F ($F = \frac{9}{5} C + 32$)." Unfortunately, this level of sophistication is rarely reached outside the physical sciences. The meaningfulness of the numerical statements we make depends on the mathematical structure of the number system we use to measure. If we use only the ordinal properties of numbers, we can only make ordinal statements, as is basically the case in a lot of standard practical use of test results.
Unless we know what we are doing when we measure pupils' ability, there will never be any real progress in this field.

Any systematic approach towards the assessment of pupils' ability implies the use of a model. The more explicit the model is, the better we know what we are doing and what the consequences are. What is a model? Without going into detail, we can say that a model is an abstract picture of a 'chunk' of reality in which relevant factors are defined and their mutual relationships are specified. A model implies the use of basic assumptions, on which the model is based. A model can have few assumptions, in which case it is usually called a weak model, or may have more assumptions, and is thus called a strong model. A weak model has relatively little explanatory power, but is less easily rejected by the emperical data, while a strong model has relatively large explanatory power, but is more easily rejected by the data. Models can also differ on the following point: some models are formulated in terms that presuppose that measurement has already taken place, while in other models measurement is a result of the theory.

The Rasch-model is a strong model where measurement is a bonus of the model itself. The Rasch-model is based on four basic ideas:

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. van Weeren

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

FL016707

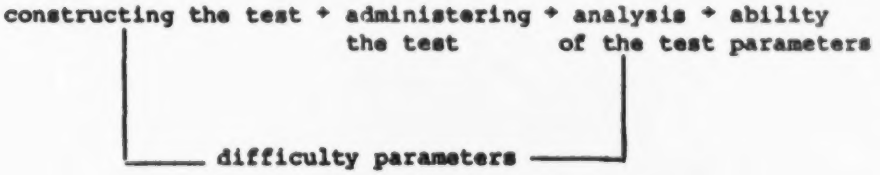
1 separation of observable behaviour and an underlying latent trait, responsible for this behaviour
2 observed behaviour is essentially probabilistic in character
3 Local stochastic independance. In terms of testing this means that the correlations between items are only caused by their relationship with the latent trait (a homogeneous test) and that in the absence of the trait, or when the trait is kept constant at a particular level, the correlation disappears.
4 Measurement of the latent trait requires the repeated assessment of indices of this trait, not necessarily the same index on each occasion.

The Rasch-model specifies that the performance of the pupil is completely determined by two variables, i.e.:
1 his ability and
2 the difficulty of the item
with the exclusion of other factors. So it excludes e.g. the time-factor, guessing behaviour and the influence of the discriminative power of the items (this is the degree of sensitivity in distinguishing between higher and lower ability). So the only things that can vary in the Rasch-model are the ability and the difficulty. We call these the parameters of the model, with symbols θ and σ. Since we take a probabilistic off round, any numerical result of the expression in which θ and σ play a role will lie in the interval $[0,1]$. It is perfectly reasonable to assume that the chance of a correct performance increases with an increase in θ and decreases with an increase in σ. Even more simply formulated one could say that the chance of correct performance increases, while $\theta - \sigma$ increases. One of the most simple statistical functions that can be used in the above situation is the logistic function.

$$P\,ij = P(+/\theta) = \frac{e^{(\theta j - \sigma i)}}{1 + e^{(\theta j - \sigma i)}}$$

The reasons for choosing this function and no other are too technical to mention here.

What are the noticeable characteristics of the Rasch-model?
1 The following sequence is appropriate to the Rasch-model:



We see here that we have a theory (of behaviour) before we have measurement. This is in contrast with classical test-theory, which assumes measurement and only then starts as a model ($X = T + E$). The origin of the measures is independent of the classical test-theory and in practice is rarely, if ever, rationalized, so that we very often do not even know the level of measurement.

2 The Rasch-model is geared towards the item-level, instead of the test-level, thereby using only qualitative information on item-level, that is, whether the item has been correctly constructed or not. Distilling quantitative information from qualitative data requires explicit theory that can be tested and, if necessary, rejected. In the Rasch-model, by quantifying the parameters, we 'explain' qualitative information. Classical test-theory uses a 'pseudo-quantitative' test-score as point of departure. The scale of measurement (the same for θ and σ) follows from the theory.

3 The Rasch-model does not try to explain in a deterministic manner why a certain answer to an item is given, but tries to explain the chance of a correct response.

4 The Rasch-model has the property of 'specific objectivity'; it is 'population-independent'. This means that comparison of ability-parameters is not dependent on a particular set of items being used and also that item-parameters can be compared, independent of the particular pupils being used to calibrate the items.
This property is common in physical measurement and is a very important property. We take it for granted, for example, that in measuring the length of a pole it is not important which yard-stick is used, neither that the outcome of the measurement is dependent on what sort of pole is being measured. It has been proven that the Rasch-model is the only model with this property.
This is all in clear contrast with classical test-theory, where item and test characteristics are really population-dependent. Classic measures like the p-values of the items and the reliability (KR20) of the test depend on the population from which the test-data were obtained
$(X = T + E)$
$(\text{Var } X = \text{var } T + \text{var } E)$

$$\text{Reli} = \frac{\text{var } T}{\text{var } X} = \frac{\text{var } T}{\text{var } T + \text{var } E} \qquad \frac{1}{1+1} = .5 \qquad \frac{3}{3+1} = .75$$

population-dependent

increase in var T in a population increases Reli, even if var E remains constant (= accuracy of measurement is constant)

5 The Rasch-model permits a clearer insight into the accuracy of ones' measurements. The Rasch-model allows a differential assessment of the accuracy of measurement, by computing the information-functions (-curves) for items and test.
This has several advantages:
a The concept of parallelism of test becomes much clearer. Two tests are defined as parallel if they measure the same ability and have the same test-information function. In this way, two or more seemingly very different tests can be made parallel.

b Using the test-information function in conjunction with the distribution of true ability of the pupils enables one to make statements about the applicability and usefulness of the tests.

Model-fit and the Rasch-model

In order to give a rough idea, one could say that, in very general terms, the assessment of model-fit could be done in the following way:

a collect observations
b estimate parameters
c use the model to make predictions about what data ought to be expected if the model were true
d find a measure to express the discrepancy between observation and expectation.

Assessing fit is still a problem. It is not really known how sensitive the existing measures of fit are to violations of what assumption(s).

Extensions of the Rasch-model

E.g.: Linear logistic model: The σ parameter can have a linear composition, expressing a theory about the difficulty of items.

An example of applications

Suppose we have a large number of items all measuring the same ability and calibrated according to the Rasch-model in a variety of different situations. It then becomes possible to recalibrate all items in such a manner that all subsequent values refer to a common scale. For a pair of tests, this could be done by the anchor item technique. However, as the number of items and their range of difficulty grows and grows, one has to think of using a network of linking tests, finally resulting in what is known as an itembank. If successfully constructed, one has assembled a very large number of items on a common scale, with a range of difficulty too large for any one single person to cope with at a given moment. From the bank, one can then generate, according to one's own specification, a large number of different tests, tailor-made, so to say. One can make easy tests, for more accurate assessment of low level ability, or difficult tests. One can make long tests and short tests, one can make tests with a wide range of difficulty or a narrow range, and so on. One can compare locally made items (e.g. on the school level) with nationally used items (e.g. Cito items), if both are in the bank. Obviously, an itembank can be used to assess development and growth, e.g. as a result of curriculum or other more sophisticated changes. A fancier application is a form of criterion-referencing whereby mastering a criterion is scored as if it were an item (1 = succes, 0 = failure) and introduced into the bank, together with other 'normal' items. Such a criterion could be passing or failing an exam, or admission, for example, into a professional training programme. The itembank can then be used to measure a person's approach to or departure from the criteria.